Journal of Water and Soil
https://jsw.um.ac.ir

Research Article
Vol. 37, No. 1, Mar.-Apr. 2023, p. 129-143

Comparison of Machine Learning Methods in the Selection of Predictors of Atmospheric-Ocean General Circulation Models

M. Amirabadizadeh[1*], M. Forozanmehr[2], M. Yaghoobzadeh[3], S. Hoseinabadi[4]

Received: 11-05-2022
Revised: 29-09-2022
Accepted: 30-11-2022
Available Online: 30-11-2022

How to cite this article:
Amirabadizade, M., Forozanmehr, M., Yaghoobzadeh, M., & Hoseinabadi, S. (2023). Comparison of Machine Learning Methods in the Selection of Predictors of Atmospheric-Ocean General Circulation Models. *Journal of Water and Soil 37*(1): 129-143. (In Persian with English abstract).
https://doi.org/10.22067/jsw.2022.76605.1166

Introduction

Nowadays, climate change is one of the human challenges in the exploitation and management of water resources. Temperature along with precipitation is one of the most important climatic elements and is one of the main factors in zoning and climatic classification. Due to location of Iran within the drought belt and proximity to the high-pressure tropical zone, this country has an arid and semi-arid climate and suffers from drought in majority of years. Therefore, temperature fluctuations and variability are important issues, and make the study of temperature changes a necessity. In the current study, four data mining algorithms in selecting predictors for downscaling of maximum temperature in Birjand synoptic station have been studied, compared and the superior algorithm has been introduced. As the number of large scale features are high, selection of machine learning algorithm will play as an important role in statistical downscaling of climatic variables such as maximum temperature.

Materials and Methods

Today, the data set is such that many variables are used to describe the climatic phenomenon in environmental studies. As the number of data is huge, choosing the predictors is one of the most important steps in preprocessing machine learning. In this study, four machine learning methods including stochastic approximation of simultaneous turbulence (SPSA), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and Gradient Boosting Method (GBM) in selecting important features in downscaling of maximum temperature in Birjand synoptic station during the statistical period of 1961-2019 were studied and compared. It is a mechanism to find a combination of predictors that with a minimum number of predictors can produce an acceptable evaluation index in estimating the variable under study. For the present study, the weather information of Birjand Synoptic Meteorological Station has been prepared by the Meteorological Organization of Iran. In order to calibrate and validate the machine learning algorithms, 70% and 30% of the available monthly data, respectively, were allocated for this purpose. To conduct this research, coding in R-Studio environment and Caret and Fscaret packages were used. In this study, to evaluate the performance of the algorithms, three indices includes relative Nash-Sutcliffe Efficiency (rNSE), Volume Efficiency (VE) and Kling-Gupta Efficiency (KGE) were used.

Results and Discussion

Before using the algorithms in selecting large-scale predictors, the correlation between these variables and the maximum observational temperature at Birjand station was investigated. Large scale variables mslp, P1_v, P8_v, P8_u, P850 Temp, with a maximum correlation temperature of 0.6 showed that the correlation is acceptable given the complexity of the climate change phenomenon. In addition, these results show that all the

1 and 3- Assistant Professor and Associate Professor, Department of Water Engineering, University of Birjand, Birjand, Iran & Members of Drought and Climate Change Research Group
(*- Corresponding Author Email: mamirabadizadeh@birjand.ac.ir)
2 and 4- Ph.D. Candidates in Water Resources, Faculty of Agriculture, University of Birjand, Birjand, Iran

algorithms used the important factors including F1, F2, F15, F16, F18, F20 and F26 by more than 50% and the first variable (mean pressure at the ocean surface) was the most important parameter in downscaling of maximum temperature. Also, the highest importance was for P1_v and the lowest value related to P5_u, as 73.2% and 15%, respectively. Violin plots of downscaled maximum temperature in validation step of different algorithms along with the observed maximum temperature in Birjand synoptic station in each of the algorithms showed that the values of the first and third quartiles in the output data of SPSA algorithm compared to other algorithms were closer to the observed data. According to the evaluation criteria, SPSA algorithm has a higher performance than other algorithms in reproducing the maximum monthly temperature values in Birjand synoptic station. Also, based on the volumetric efficiency evaluation criteria and relative Nash-Sutcliffe, GBM algorithm was more successful in selecting predictors than Ridge and LASSO algorithms. It is also observed that SPSA algorithm shows different results than other algorithms. In comparison of mean and variance of downscaled and observed maximum temperature, the results of t-test and F-test showed that SPSA algorithm has higher efficiency than other algorithms in regenerating mean and variance of observed maximum temperature in Birjand synoptic station at the 5% significance level.

Conclusion

The data used in this study included large scale atmospheric variables and the maximum observed temperature at Birjand station. The algorithms were used to select important predictors and the performance of these methods in the validation part. According to the results of this study, the highest importance among large-scale variables is related to P1_v and the lowest value is related to P5_u, the values of which were 73.2% and 15%, respectively. The SPSA algorithm also performs better than other algorithms in selecting predictors and consequently the maximum temperature.

Keywords: Atmosphere- Ocean general circulation model, Downscaling, Feature selection, Machine learning algorithm, Maximum temperature

نشریه آب و خاک
https://jsw.um.ac.ir

مقاله پژوهشی

جلد ۳۷، شماره ۱، فروردین- اردیبهشت ۱۴۰۲، ص. ۱۴۳-۱۲۹

مقایسه روش‌های یادگیری ماشین در انتخاب پیش‌بینی‌کننده‌های مدل‌های گردش عمومی جو- اقیانوسی برای ریزمقیاس‌نمائی دمای بیشینه (مطالعه موردی: بیرجند)

مهدی امیرآبادی‌زاده[1*]- مهدیه فروزان‌مهر[2]- مصطفی یعقوب‌زاده[3]- سعیده حسین‌آبادی[4]

تاریخ دریافت: ۱۴۰۱/۰۲/۲۱
تاریخ بازنگری: ۱۴۰۱/۰۷/۰۷
تاریخ پذیرش: ۱۴۰۱/۰۹/۰۹

چکیده

افزایش غلظت گازهای گلخانه‌ای در اتمسفر باعث تغییرات زیادی در مؤلفه‌های اقلیمی کره زمین شده است که این تغییرات در پارامترهای اقلیمی به صورت افزایشی یا کاهشی است. امروزه تغییر اقلیم یکی از چالش‌های بشر در بهره‌برداری و مدیریت منابع آب است، همچنین شرایط کنونی آب و هوای جهانی نشان دهنده افزایش خطرات ناشی از پدیده خشکی در بسیاری از مناطق جهان در آینده است. مدل‌های گردش کلی جوی یکی از مهم‌ترین و پرکاربردترین روش‌ها در مطالعات تغییرات اقلیمی در مقیاس منطقه‌ای می‌باشد. یکی از اولویت‌های اصلی ریزمقیاس‌نمایی آماری انتخاب پیش‌بینی‌کننده‌ها به عنوان ورودی به مدل ریزمقیاس‌نمایی در پژوهش می‌باشد. برای انتخاب پیش‌بینی‌کننده‌های مهم از بین ۲۶ متغیر جو بالا، از چهار الگوریتم یادگیری ماشین شامل لاسو، ستیغی، GBM، SPSA در ریزمقیاس‌نمایی آماری دمای بیشینه در ایستگاه بیرجند استفاده گردید و عملکرد این روش‌ها با سه شاخص نش-ساتکلیف نسبی، کلینگ-گوپتا و بازده حجمی در بخش صحت‌سنجی مورد بررسی قرار گرفت. نتایج نشان داد که بیشترین میزان اهمیت برای مؤلفه سرعت نصف‌النهاری نزدیک سطح و کمترین مقدار آن در مؤلفه سرعت مداری در ارتفاع ۵۰۰ هکتو پاسکال می‌باشد که مقادیر آن به‌ترتیب ۷۳/۲٪ و ۱۵٪ تعیین شد. همچنین نتایج شاخص‌های ارزیابی عملکرد نش- ساتکلیف نسبی و کلینگ-گوپتا، نشان دادند که الگوریتم SPSA دارای عملکرد بهتری از سایر الگوریتم‌ها درانتخاب پیش‌بینی‌کننده‌ها و به تبع آن ریزمقیاس‌نمائی دمای بیشینه می‌باشد. مقایسه میانگین و واریانس خروجی ریزمقیاس شده توسط الگوریتم‌های مورد استفاده و داده‌های مشاهداتی در بخش صحت‌سنجی نشان داد که الگوریتم SPSA نسبت به سایر الگوریتم‌ها در باز تولید میانگین و واریانس دمای بیشینه مشاهداتی در ایستگاه سینوپتیک بیرجند دارای توانایی بیشتری می‌باشد.

واژه‌های کلیدی: الگوریتم یادگیری ماشین، انتخاب پیش‌بینی‌کننده، دمای بیشینه، ریزمقیاس‌نمائی، مدل گردش عمومی جو- اقیانوسی

مقدمه

امروزه تغییر اقلیم یکی از چالش‌های بشر در بهره‌برداری و مدیریت منابع آب است، تحقیقاتی که در این زمینه در ایران و سایر کشورها صورت گرفته است، نشان دهنده تغییر در پارامترهای اقلیمی به صورت افزایشی یا کاهشی است. دما در کنار بارش از مهم‌ترین عناصر اقلیمی محسوب می‌شود که در تعیین نقش و پراکندگی دیگر عناصر اقلیمی نیز مؤثر است. همچنین دما از عوامل اصلی و اساسی در پهنه‌بندی و طبقه‌بندی اقلیمی محسوب می‌شود، بر این اساس نوسانات و تغییرپذیری دما دارای اهمیت زیادی است (Sfandiari *et al.*, 2010). درباره تغییر اقلیم و بررسی عناصر هواشناسی مطالعات متعددی در سطح ایران و جهان صورت گرفته است.

به‌طور کلی تحقیقاتی که در زمینه تغییر اقلیم در جهان صورت گرفته است، بیان‌کننده‌ی این حقیقت است که تغییر هرچند کم در دما

۱ و ۳- به‌ترتیب استادیار و دانشیار گروه علوم و مهندسی آب، اعضای گروه پژوهشی خشکسالی و تغییر اقلیم، دانشگاه بیرجند، بیرجند، ایران
(*- نویسنده مسئول: Email: mamirabadizadeh@birjand.ac.ir)
۲ و ۴- دانشجویان دکتری منابع آب، دانشکده کشاورزی، دانشگاه بیرجند، بیرجند، ایران

DOI: 10.22067/jsw.2022.76605.1166

موجب تغییر در وقوع پدیده‌هایی نظیر خشکسالی، بارش‌های سنگین و طوفان می‌شود (Balling and Idso, 1990). برای بررسی رفتار داده‌های آماری، انتخاب، شناسایی و مدل‌سازی و نیز برای کشف نظم و یا رابطه‌ای که در ابتدا ناشناخته است می‌توان از الگوریتم‌های یادگیری ماشین استفاده کرد. در ادامه چند نمونه از تحقیقات موجود در زمینه داده کاوی ارائه شده است.

نیک‌نام (Niknam, 2013) طی تحقیقی با استفاده از ۲۰ متغیر اقلیمی طی دوره ۱۰ ساله(۱۹۹۵-۲۰۰۵) مربوط به ایستگاه‌های اقلیمی و سینوپتیک استان اصفهان و با استفاده از الگوریتم‌های داده‌کاوی درخت‌های تصمیم‌گیری، شبکه بیزین و شبکه عصبی به داده‌کاوی اقلیمی جهت ارائه مدل پیش‌بینی اقلیم در این استان پرداخته است، در پایان نیز عوامل تأثیرگذار بر پهنه‌بندی اقلیمی استان توسط مدل پیشنهادی ارائه شده است. در تحقیقی صلاحی و فاطمی‌نیا (Salahi and Fateminiya, 2017)، به منظور ایجاد داده‌های روزانه تا سال ۲۰۳۰ برای پیش بینی روند تغییرات یخبندان در ایستگاه هواشناسی کاشان، از خروجی پنج مدل گردش کلی شامل BCM2، HADCM3، IPCM4، GLAOM و HADGEM استفاده کردند که درنهایت مدل گردش کلی GLAOM به منظور پیش‌بینی داده‌های مصنوعی دمای کمینه و بیشینه و همچنین بررسی یخبندان‌های کاشان در بازه اقلیمی آینده برگزیده شد. طی تحقیقی فاتحی و همکاران(Fatahi *et al.*, 2012)، با استفاده از شاخص بارندگی و روش قواعد انجمنی به پیش‌بینی خشکسالی پرداختند، نتایج حاصل از این تحقیق نشان می‌دهد الگوریتم تهیه شده قادر است شرایط خشکی متوسط تا رطوبت شدید و همچنین شرایط خنثی را با احتمال بسیار زیاد و اطمینان بالا پیش‌بینی نماید.

با توجه به پیچیدگی پدیده تغییر اقلیم و نیز تعداد بالای متغیرهای پیش‌بینی‌کننده خروجی مدل‌های گردش کلی(در این تحقیق ۲۶ متغیر جو بالا)، انتخاب پیش‌بینی‌کننده‌های با اهمیت‌تر و حذف متغیرهای کم اهمیت می تواند روشی راهگشا در ریزمقیاس نمائی خروجی مدل‌های گردش کلی باشد. برای تشخیص و انتخاب پیش‌بینی‌کننده از همبستگی متغیرهای پیش‌بینی‌کننده –پیش‌بینی شونده نیز استفاده شده است. چن و همکاران (Chen *et al.*, 2012) و مینو و همکاران (Meenu *et al.*, 2013) در انتخاب پیش‌بینی‌کننده‌ها در ریزمقیاس‌نمائی از آنالیز همبستگی بین پیش‌بینی‌کننده-پیش‌بینی‌شونده استفاده نمودند. ناصری و زهرایی (Nasseri and Zahraie, 2013) از روش آنالیز مؤلفه اصلی برای تشخیص پیش بینی‌کننده‌های با اهمیت برای ریزمقیاس‌نمایی با روش SDSM استفاده کردند. نتایج مطالعه آن‌ها نشان داد که رطوبت نسبی و سرعت مهم‌ترین پیش‌بینی‌کننده‌ها در این تحقیق بودند.

بررسی منابع در روش‌های انتخاب ویژگی، نشان می‌دهد که روش رگرسیون گام به گام یکی از کاربردی‌ترین روش‌ها است، اما دارای کاستی‌هایی است. این روش زمانی که تعداد زیادی پیش‌بینی کننده وجود داشته باشد که همبستگی قوی با هم دارند دارای نقص است. حسامی و همکاران (Hessami *et al.*, 2008) از روش آنالیزرگرسیون گام به گام برای انتخاب بهترین پیش‌بینی‌کننده‌ها برای ریزمقیاس‌نمائی با SDSM برای شبیه‌سازی مقادیر حدی استفاده کردند.

تعداد کمی از مطالعات ریزمقیاس‌سازی از الگوریتم LASSO برای انتخاب پیش‌بینی‌کننده در گذشته استفاده کردند. اما، کاربرد آن در سال‌های اخیر، رو به گسترش است. هی و همکاران (He *et al.*, 2019) ازالگوریتم LASSO برای تعیین بهترین پیش‌بینی‌کننده استفاده کردند و ارتفاع ژئوپتانسیل در ۵۰۰ هکتوپاسکال (hPa) برای ریزمقیاس‌نمایی بارندگی ماهانه در دره رودخانه یانگ تسه در چین بعنوان مهم‌ترین پیش‌بینی‌کننده معرفی گردید. عملکرد سه روش رگرسیون LASSO، OLS و ستیغی توسط متوکریشنان و روحینی (Muthukrishnan and Rohini, 2016) در برازش مدل برای تحلیل سلامت افراد با استفاده از داده های پایگاه اندازه‌گیری مشخصات بیماران دیابتی مورد بررسی قرار گرفت. در این تحقیق از ۴۲۲ رکورد مشاهداتی استفاده گردید. نتایج این تحقیق نتایج نشان داد که روش LASSO با کوچک کردن ضرایب مدل به سمت مقدار صفر، بهتر از روش‌های دیگر مورد مطالعه در تحقیق عمل می‌کند. از این رو، روش LASSO را می‌توان به عنوان جایگزینی برای روش های انتخاب ویژگی مرسوم استفاده کرد. آنها تأکید نمودند که نتیجه تحقیق برای محققینی که در وظایف یادگیری ماشینی کار می‌کنند، مفید خواهد بود.

ژانگ و همکاران (Zhang *et al.*, 2016) از BMA[1] به عنوان یک روش ریزمقیاس‌سازی برای تولید دمای میانگین منطقه‌ای استفاده کردند. در انتخاب پیش‌بینی‌کننده دراین روش، برتری متغیرها بر اساس قضیه بیز تشخیص داده می‌شود و بر اساس احتمال مشخص می‌شود. از این رو، وزن‌های تخمینی برای پیش‌بینی‌کننده مستقیماً دقت آن متغیر را منعکس می‌کنند و به نتایج واقعی‌تری می‌رسند.

پناهی و میرهاشمی(Panahi and Mirshahi, 2016)، با استفاده از دو الگوریتم یادگیری ماشین CART و CHAID به پیش‌بینی دمای‌ها در ایستگاه سینوپتیک اراک پرداخته‌اند. نتایج حاصل از این تحقیق نشان داد که مدل درختی CHAID در مقایسه با مدل CARTعملکرد بهتری در پیش‌بینی متوسط دمای ماهانه ماه بعد دارد، بنابراین می‌توان از این مدل برای پرکردن داده‌های هواشناسی گم شده استفاده کرد. ترانسکو و همکاران (Troncoso *et al.*,

1- Bayesian Model Averaging

2015)، در تحقیقی سرعت باد را در مناطق بادخیز با استفاده از الگوریتم RTs پیش‌بینی نموده‌اند. در پژوهشی دیگر با استفاده از داده‌های هواشناسی مؤسسه تحقیقات کشاورزی هند طی دوره ۴ ساله (۲۰۰۴–۲۰۰۷) شامل دمای بیشینه و کمینه، ساعات آفتابی، سرعت باد، رطوبت نسبی به مقایسه مقادیر تبخیر روزانه تشت و تخمین‌های شبکه عصبی، رگرسیون و مدل‌های اقلیمی پرداخته‌اند (Diamantopoulu *et al.*, 2010).

پال و دسوال (Pal and Deswal, 2009)، برای مدل‌سازی تبخیر و تعرق در ایستگاه دیویس در کالیفرنیا از مدل درختی M5P استفاده کردند و پتانسیل این روش را با روش‌های پنمن- مونتیث، فائو-۵۶ و هارگریوز- سامانی کالیبره و مقایسه کردند، طبق نتایج به‌دست آمده در این تحقیق می‌توان از مدل M5P به‌عنوان مدل مرجع در تعیین تبخیر و تعرق استفاده نمود. همچنین در سال ۱۳۹۳، امیدوار و همکاران (Omidvar *et al.*, 2015)، طی تحقیقی به ارزیابی کارایی مدل درخت تصمیم در پیش‌بینی بارش ایستگاه سینوپتیک کرمانشاه پرداختند، در این تحقیق از الگوریتم CART به‌عنوان یکی از انواع درختان تصمیم رگرسیونی جهت پیش‌بینی بارش ۳۰ ماه بعد استفاده شد.

جعفرزاده و همکاران (Jafarzadeh *et al.*, 2021) برای ریزمقیاس‌نمائی بارش در ایستگاه سینوپتیک بیرجند و ارومیه از روش SVM استفاده نمودند و ۲۶ پارامتر NCEP/NCAR بعنوان ورودی-های ریزمقیاس‌نمایی مورد استفاده قرار گرفتند. آن‌ها پنج پیش‌بینی کننده برتر را با روش‌های PCA، رگرسیون گام به گام، آنالیز همبستگی و BMA انتخاب و نتایج ریزمقیاس‌نمایی بارش روزانه را مورد ارزیابی قرار دادند. نتایج این پژوهش نشان داد که روش‌های BMA و رگرسیون گام به گام دارای توانمندی بالاتری در انتخاب پیش‌بینی‌کننده برای ریزمقیاس‌نمایی بارش در این ایستگاه‌ها می باشند.

الگین و همکاران (Algin *et al.*, 2022) در پژوهشی از تقریب تصادفی آشفتگی همزمان (SPSA)، بهینه‌سازی پرندگان مهاجر و الگوریتم تبرید شبیه‌سازی شده برای تولید زیر مجموعه‌ها استفاده نمودند. زیر مجموعه‌های تولید شده توسط الگوریتم‌ها، با استفاده از انتخاب ویژگی مبتنی بر همبستگی ارزیابی شده و عملکرد آنها با استفاده از درخت تصمیم (C4.5) به عنوان طبقه‌بندی کننده مورد سنجش قرارگرفت. آن‌ها برای اولین بار، الگوریتم SPSA را به عنوان یک رویکرد فیلتر برای انتخاب ویژگی مورد استفاده قرار دادند. همچنین نتایج نشان داد که الگوریتم SPSA از نظر مقادیر دقت بهتر از سایر الگوریتم‌ها عمل می‌کند. نکته دیگر این است که همه الگوریتم‌ها تعداد ویژگی‌ها را بیش از ۵۰ درصد کاهش دادند.

مرور منابع نشان‌دهنده محدود بودن استفاده از تکنیک‌های یادگیری ماشین در انتخاب ویژگی در ریزمقیاس‌سازی متغیرهای اقلیمی مانند دمای بیشینه هوا بوده و اینکه هنوز کاربرد آن‌ها گسترش نیافته است. بنابراین در پژوهش حاضر، الگوریتم‌های تقریب استوکاستیک آشفتگی همزمان، ستیغی، لاسو و ماشین تقویت شده گرادیان برای انتخاب پیش‌بینی‌کننده در ریزمقیاس‌سازی دمای بیشینه در ایستگاه سینوپتیک بیرجند مورد بررسی و مقایسه قرار گرفت و الگوریتم برتر معرفی شد.

مواد و روش‌ها

امروزه مجموعه داده‌ها بگونه‌ای است که در توصیف پدیده مورد بررسی از متغیرهای زیادی بهره می‌گیرد. انتخاب پیش‌بینی‌کننده[1] یکی از مهم‌ترین گام‌ها در پیش پردازش در یادگیری ماشین می‌باشد. بعبارت دیگر مکانیزمی به منظور یافتن ترکیبی از پیش‌بینی‌کننده‌ها می‌باشد که با حداقل تعداد پیش‌بینی‌کننده بتواند در برآورد متغیر مورد بررسی شاخص ارزیابی قابل قبولی را تولید نماید. از جمله مزایای انتخاب پیش‌بینی‌کننده می‌توان به کاهش زمان پیش پردازش، ساده-سازی مدل‌ها وزمان تحلیل کوتاه‌تر اشاره نمود. در این تحقیق هدف مقایسه چند الگوریتم در ریزمقیاس‌سازی دمای بیشینه ایستگاه همدیدی بیرجند طی دوره آماری ۱۹۶۱-۲۰۱۹ می‌باشد که از ۲۶ پارامتر مرتبط با جو بالای مستقل به‌عنوان پیش‌بینی‌کننده استفاده شد (پیش‌بینی‌کننده‌ها با حرف F در اشکال و جداول نشان داده شده اند). توصیف این پیش‌بینی‌کننده‌های جو بالا در جدول ۱ آورده شده است.

منطقه مورد مطالعه

برای انجام پژوهش حاضر اطلاعات آب و هوایی ایستگاه هواشناسی سینوپتیک بیرجند از سازمان هواشناسی کشور تهیه شده است. شهرستان بیرجند واقع در استان خراسان جنوبی و مرکز این استان می‌باشد. این شهرستان از شمال به شهرستان قائنات، از شمال شرق به شهرستان زیرکوه، از شرق به شهرستان‌های درمیان و سربیشه،از جنوب به شهرستان خوسف و از غرب به شهرستان سرایان محدود است. مختصات جغرافیایی این شهرستان ۵۹ درجه و ۱۳ دقیقه طول جغرافیایی و ۳۲ درجه و ۵۳ دقیقه عرض جغرافیایی می باشد، همچنین این شهرستان در ارتفاع ۱۴۷۰ متری از سطح دریا قرار گرفته و دارای مساحت ۳۹۴۹ کیلومتر مربعی است. به‌طور کلی آب و هوای استان خراسان جنوبی خشک و بیابانی است، شهرستان بیرجند به دلیل دوری از مسیر بادهای باران‌آور، عدم وجود ارتفاعات مهمی که بتواند در تعدیل آب و هوا و جذب نزولات جوی مؤثر واقع شود و همچنین مجاورت با کویر لوت در خشک ترین تقسیم‌بندی اقلیمی خراسان قرار دارد.

1- Feature Selection

جدول ۱- تعاریف ۲٦ متغیر بزرگ‌مقیاس NCEP (hpa مخفف هکتو پاسکال است)

Table 1- Definition of 26 large scale NCEP variable (hpa is abbreviation for Hectopascal)

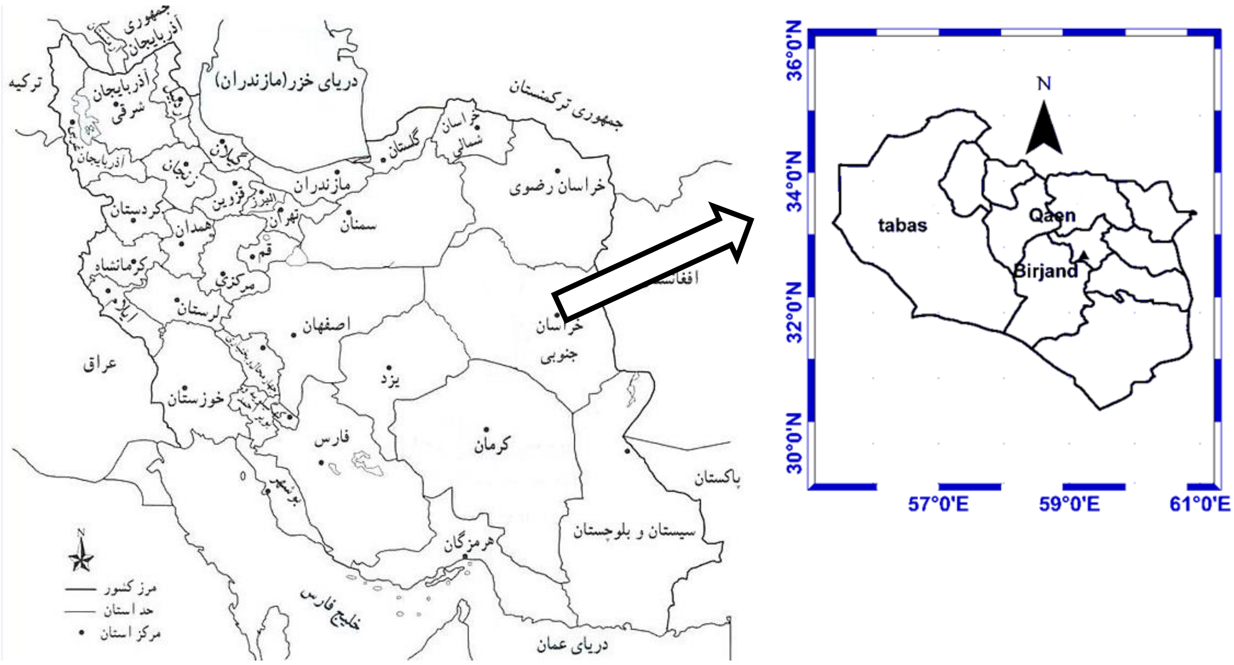
ویژگی Property	متغیربزرگ مقیاس Large scale variable	واحد Unit	ویژگی Property	متغیربزرگ مقیاس Large scale variable	واحد Unit
F14	نیروی جریان هوا در ارتفاع 850hpa (p8_f) 850 h Pa Wind speed	m/s	F1	میانگین فشار سطح دریا (mslp) Mean sea level pressure	Pa
F15	مؤلفه سرعت مداری در ارتفاع 850hpa (p8_u) 850 h Pa Zonal velocity	m/s	F2	سرعت جریان هوا ژئوستروفیک در نزدیک سطح (p1_f) 1000 h Pa Wind speed	m/s
F16	مؤلفه سرعت نصف النهاری در ارتفاع 850hpa (p8_v) 850 h Pa Meridional velocity	m/s	F3	مؤلفه سرعت مداری نزدیک سطح (p1_u) 1000 h Pa Zonal velocity	m/s
F17	تاوایی در ارتفاع 850hpa (p8_z) 850 h Pa Vorticity	s^{-1}	F4	مؤلفه سرعت نصف النهاری نزدیک سطح (p1_v) 1000 h Pa Meridional velocity	m/s
F18	جهت باد در ارتفاع hPa۸۵۰ (p8th) 850 h Pa Wind direction	Deg	F5	تاوایی نزدیک سطح (p1_z) 1000 h Pa Vorticity	s^{-1}
F19	واگرایی در ارتفاع 850hpa (p8zh) 850 h Pa Divergence	s^{-1}	F6	واگرایی نزدیک سطح (p1_th) 1000 h Pa Wind direction	s^{-1}
F1\20	ارتفاع ژئوپتانسیل سطح 500hpa (p500) 500 h Pa Geopotential height	M	F7	جهت باد نزدیک سطح (p1zh) 1000 h Pa Divergence	Deg
F21	ارتفاع ژئوپتانسیل سطح 850hpa (p850) 850 h Pa Geopotential height	M	F8	نیروی جریان هوا در ارتفاع 500hpa (p5_f) 500 h Pa Wind speed	Pa
F22	رطوبت نسبی در نزدیک سطح (prcp) Precipitation	%	F9	مؤلفه سرعت مداری در ارتفاع 500hpa (p5_u) 500 h Pa Zonal velocity	m/s
F23	رطوبت نسبی در ارتفاع 500hpa (s500) 500 h Pa Specific humidity	%	F10	مؤلفه سرعت نصف النهاری در ارتفاع 500hpa (p5_v) 500 h Pa Meridional velocity	m/s
F24	رطوبت نسبی در ارتفاع 850hpa (s850) 850 h Pa Specific humidity	%	F11	تاوایی در ارتفاع 500hpa (P5_z) 500 h Pa Vorticity	s^{-1}
F25	رطوبت ویژه در نزدیک سطح (shum) 1000 h Pa Specific humidity	Kg/kg	F12	جهت باد در ارتفاع 500hpa (p5th) 500 h Pa Wind direction	Deg
F26	میانگین دما در ارتفاع ۲ متری (temp) Screen(2 m) air temperature	K	F13	واگرایی در ارتفاع 500hpa (p5zh) 500 h Pa Divergence	s^{-1}

حداکثر دمای مطلق در این شهرستان ۴۴ درجه سانتی‌گراد می باشد (Skhmet.ir).

انتخاب پیش‌بینی‌کننده‌ها (ویژگی‌ها)

در این تحقیق از چهار روش شامل تقریب استوکستیک آشفتگی همزمان، لاسو[1]،ستیغی[2]و روش GBM[3] در انتخاب ویژگی‌های مهم بصورت یکپارچه[4] در ریزمقیاس‌نمایی دمای بیشینه مورد بررسی و مقایسه قرار گرفت. به‌منظور واسنجی و صحت‌سنجی الگوریتم‌ها، بترتیب ۷۰ و ۳۰ درصد داده‌های در اختیار، به این منظور اختصاص یافت. برای اجرای این تحقیق از کد نویسی در محیط نرم افزاری RStudio و بسته‌های Caret و Fscaret استفاده گردید. بمنظور اجرای این الگوریتم‌ها، دو بردار ورودی و خروجی تشکیل گردید که بردار ورودی مشتمل بر ۲۶ ویژگی مدل بزرگ مقیاس گردش کلی جو و بردار خروجی آن دمای بیشینه مشاهداتی طی دوره زمانی مورد پژوهش می‌باشد.

1- Least Absolute Shrinkage and Selection Operator (LASSO)
2- Ridge
3- Gradient Boosting Machine
4- Ensemble



شکل ۱- موقعیت جغرافیایی شهرستان بیرجند
Figure 1- Geographical location of Birjand city

تقریب استوکاستیک آشفتگی همزمان[1] (SPSA)

این الگوریتم بهینه‌سازی تکراری به طور کلی در مسایل غیر خطی استفاده می‌شود که متغیرهای زیادی دارند و در آن‌ها برآورد شیب تابع هدف دشوار یا غیر ممکن است. کد مورد استفاده در این روش، شکلی از توابع تصادفی است که از ایده تقریب تصادفی برای تخمین گرادیان عملکرد استفاده می‌کند و آن‌ها را به یک الگوریتم شدیدترین کاهش می‌رساند.

این روش، از الگوریتم بهینه‌سازی گرادیان کاهشی استفاده می نماید که با مقادیر اولیه $\widehat{\theta_k}$ شروع شده و با لحاظ رابطه زیر مقادیر حدس های بعدی در جهت رسیدن به جواب را دنبال می نماید:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \widehat{g_k}(\hat{\theta}_k) \qquad (۱)$$

که در آن ak یک توالی افزایشی تکرارهای غیر منفی، $\widehat{g_k}(\hat{\theta}_k)$ بیانگر شیب تقریبی در $\hat{\theta}_k$ می‌باشد. در این الگوریتم مقدار $\widehat{g_k}$ با توجه به ارزیابی توابع جریمه در هر مرحله محاسبه می‌شود. این الگوریتم مقدار جاری w را با مقدار کمی در هر جهت به صورت w + δ و w - δ که δ > 0 تکرار می‌کند، آشفته می‌نماید. مقدار آشفتگی δ به عنوان $C_k \times k$ در نظر گرفته می‌شود و C_k یک دنباله افزایشی گرادیان غیر منفی و k بردار اغتشاش همزمان p-بعدی است (Aksakalli and Malekipirbazari, 2016).

الگوریتم ستیغی[2] (Ridge)

الگوریتم ستیغی برای انتخاب ویژگی‌های منتخب از مجموعه‌ی ویژگی‌های در اختیار، می‌تواند مورد استفاده قرار گیرد و از شکل اصلاح شده‌ی تابع جریمه رگرسیون خطی استفاده می‌نماید که در معادله زیر بیان شده است :

$$L_1 = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_{obs}^i - y_{pred}^i)^2 + \propto \sum_{j=1}^{n} a_j^2 \qquad (۲)$$

که در آن a_j ضریب فاکتور j-ام، y_{obs}^i مقدار مشاهداتی، y_{pred}^i مقدار تولید شده توسط مدل رگرسیون، $\propto$ یک فراپارامتر است که شدت جریمه جمله انتهایی را تنظیم می‌کند ومقدار آن با استفاده از اعتبارسنجی متقابل[3] تعیین می‌شود، N_t تعداد داده برای آموزش مدل و n تعداد ویژگی مورد ارزیابی می‌باشد. ایده پشت رگرسیون ریج، کوچک کردن مقادیر ضرایب به منظور حذف کردن ویژگی‌های

1- Simultaneous Perturbation Stochastic Approximation

2- Ridge

3- Cross Validation

غیرمفید است. اگر α صفر باشد، رفتار شبیه رگرسیون خطی خواهد بود. اگر α خیلی بزرگ باشد، قسمت اول تابع هزینه نادیده گرفته می‌شود و نتایج غیر قابل اعتماد هستند. بنابراین، برای اینکه مدل به درستی کار کند، نیاز به تنظیم این هایپرپارامتر وجود دارد.

روش لاسو (LASSO)[1]

در این الگوریتم که یک روش اتوماتیک قدرتمند در یافتن ویژگی‌های مهم می‌باشد، تابع هزینه براساس معادله زیر که تغییر شکل یافته الگوریتم رگرسیون خطی می باشد، مورد استفاده قرار می‌گیرد:

$$L_1 = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_{obs}^i - y_{pred}^i)^2 + \propto \sum_{j=1}^{n}|a_j| \quad (۳)$$

که پارامترهای این تابع هزینه در بخش مربوط به الگوریتم ستیغی معرفی گردید. اگر دو ویژگی به صورت خطی همبستگی داشته باشند، حضور همزمان آن‌ها مقدار تابع جریمه را افزایش می‌دهد، بنابراین الگوریتم سعی می‌کند ضریب ویژگی کم اهمیت را به صفر نزدیک کند تا بهترین ویژگی‌ها را انتخاب نماید.

روش GBM[2]

ماشین تقویت‌شده گرادیان (GBM) یک الگوریتم یادگیری ماشینی بسیار محبوب هستند که در بسیاری از حوزه‌ها موفق بوده و یکی از روش‌های کارآمد درمدل‌سازی است. در این روش از یادگیری گروهی رو به جلو برای آموزش مدل استفاده می‌شود. در مقایسه با روش‌های محبوب دیگر مانند جنگل تصادفی می‌توان گفت که در حالی که جنگل‌های تصادفی مجموعه‌ای از درختان مستقل عمیق می‌سازند، GBMها مجموعه‌ای از درختان کم‌عمق و ضعیف متوالی را با یادگیری و بهبود هر درخت نسبت به درخت قبلی ایجاد می‌کند. تابع هدف در این روش عبارتست از:

$$F(x) = argmin_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad (۴)$$

که آن L تابع جریمه و y_i مقدار مشاهداتی و γ مقدار پیش بینی شده می‌باشد. تابع جریمه میانگین مربعات خطا می باشد:

$$L = \frac{1}{n}\sum_{i=}^{n}(y_i - \gamma_i)^2 \quad (۵)$$

که پارامترهای رابطه (۵) قبلا معرفی شده‌اند (Natekin and Knoll, 2013).

شاخص‌های ارزیابی

برای ارزیابی عملکرد الگوریتم‌های مورد بررسی در این مطالعه در انتخاب پیش‌بینی‌کننده از سه شاخص عملکرد نش-ساتکلیف نسبی (rNSE)، بازده حجمی (VE[3]) و شاخص عملکرد کلینگ-گوپتا[4](KGE) استفاده گردید که روابط مورد استفاده در هر شاخص در ادامه ذکر می‌گردد. از جمله خصوصیات شاخصKGE اینست که از سه عامل ضریب همبستگی خطی بین جریان‌های مشاهده شده و شبیه‌سازی شده (r)، نسبت انحراف استاندارد بین جریان‌های مشاهده شده و شبیه‌سازی شده (α) و خطای وزنی بین جریان مشاهده شده و شبیه‌سازی شده (β) استفاده می‌نماید. بنابراین ارزیابی با این شاخص که بین ∞- تا یک تغییر می‌نماید، دارای مزیت سه‌گانه در مقایسه مقادیر مشاهداتی و شبیه‌سازی شده می‌باشد. روابط مورد استفاده در محاسبه این شاخص‌ها طی معادلات ۶ تا ۸ در ادامه ارائه شده‌اند:

$$rNSE = 1 - \frac{\sum_{i=1}^{n}(\frac{y_{i\,sim} - y_{i\,obs}}{y_{i\,obs}})^2}{\sum_{i=1}^{n}(\frac{y_{i\,obs} - \overline{y_{obs}}}{\overline{y_{1\,obs}}})^2} \quad (۶)$$

$$VE = 1 - \frac{\sum_{i=1}^{n}|y_{i\,sim} - y_{i\,obs}|}{\sum_{i=1}^{n} y_{i\,obs}} \quad (۷)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2} \quad (۸)$$

که در این روابط: $y_{i\,obs}$ بیانگر درجه حرارت بیشینه مشاهداتی، $y_{i\,sim}$ درجه حرارت بیشینه شبیه سازی شده، $\overline{y_{1\,obs}}$ میانگین دمای بیشینه مشاهداتی، r ضریب همبستگی خطی بین مقادیر دمای بیشینه مشاهداتی و شبیه سازی شده، β نسبت میانگین دمای بیشینه شبیه سازی شده به میانگین دمای بیشینه مشاهداتی و α نسبت انحراف معیار دمای بیشینه شبیه‌سازی شده به انحراف معیار دمای بیشینه مشاهداتی می‌باشد.

نتایج و بحث

در این تحقیق از چهار الگوریتم که در بخش مواد و روش‌ها تشریح گردید، برای انتخاب پیش‌بینی‌کننده‌های مهم در ریزمقیاس نمایی دمای بیشینه در ایستگاه دیده‌بانی بیرجند استفاده گردید و بر اساس شاخص‌های ارزیابی آماری مورد مقایسه قرار گرفت.

ارتباط بین ۲۶ متغیر جو بالا و متغیر پیش بینی شونده (دمای بیشینه) در شکل ۲ و شکل۳ نشان داده شده است. همانگونه که مشاهده می‌شود دو پیش‌بینی‌کننده mslp(F1) و Temp(F26) دارای بالاترین ضرایب همبستگی و P8_z (F17) دارای کمترین میزان همبستگی با درجه حرارت بیشینه در ایستگاه دیدبانی بیرجند می

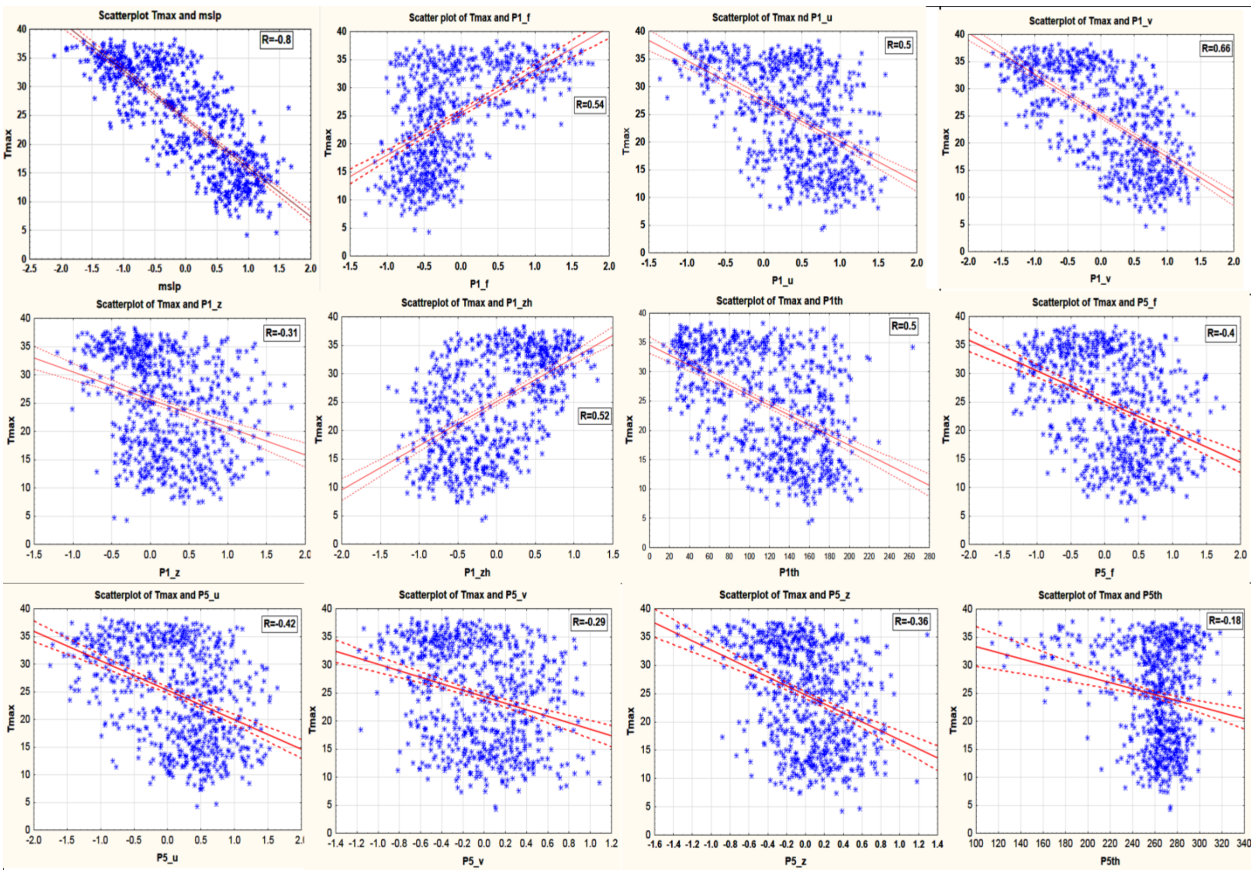
1- Least Absolute Shrinkage and Selection Operator
2- Gradient Boosting Machine
3- Volumetric Efficiency
4- Kling-Gupta Efficiency

باشند.

همچنین مشاهده می‌شود که متغیرهـای mslp, P1_v, P8_v, Temp, P8_u, P850 بـا دمـای بیشـینه دارای همبسـتگی بزرگتـر مساوی ۰/۶ می‌باشند که با توجـه بـه پیچیـدگی پدیـده تغییـر اقلـیم همبستگی قابل قبولی می‌باشد (Nazeri Tahroudi *et al.*, 2017). لازم بذکر است که بیشترین همبستگی مقدار ۰/۸۸- می‌باشد که بین میانگین فشار در سطح اقیانوس و دمای بیشینه مشاهـده شـده اسـت. انتظار می‌رود که الگوریتم‌های مورد استفاده دراین تحقیق ایـن پـیش بینی‌کننده‌ها را در لیست کاندیدای ورودی خود پیشنهاد نمایند.

همچنین با توجه به پیچیده بودن پدیده تغییر اقلـیم، همبسـتگی غیر خطی بین دمای بیشینه مشاهداتی و ۲۶ متغیر پیش‌بینی‌کننده جو بالا، مورد بررسی و تحلیل قرار گرفت. نتایج این بررسـی در جـدول ۲ نشان داده شده است. همانگونه که در این جدول مشـاهده مـی‌شـود، حداکثر همبستگی غیر خطی بین متغیر mslp (F1) و دمـای بیشـینه مشاهداتی وجود دارد. همچنین متغیـر temp (F26) در رتبـه بعـدی بالابودن همبستگی غیرخطی با دمای بیشینه مشاهداتی قـرار دارد. از سوی دیگر کمترین مقدار همبستگی غیرخطـی بـین P8_z (F17) و دمای بیشینه مشاهداتی در ایستگاه همدیدی بیرجند وجود دارد.

اهمیت هر یک از ۲۶ پارامتر پیش‌بینی‌کننده برای انتخاب بعنوان ورودی در ریز مقیاس‌نمایی دمای بیشینه با استفاده از الگـوریتم‌هـای چهارگانه مورد مطالعه در ایـن تحقیـق در شـکل ۴ نشـان داده شـده است. همانگونه که در این شکل دیده می‌شـود، همـه الگـوریتم‌هـای مورد استفاده اهمیت فاکتورهـای F1، F2، F15، F16، F18، F20 و F26 را بیشتر از ۵۰ درصد نشان می‌دهد. همانگونه که بر اساس آنالیز همبستگی انتظار می‌رفت، متغیر اول (فشار در سطح اقیانوس) بعنـوان مهم‌ترین پارامتر در ریزمقیاس‌نمایی، پیشنهاد شـده اسـت. در مقابـل متغیرهای پنجم، هفتم، هشتم، دوازدهم، بیسـت و چهـارم و بیسـت و پنجم در همه الگـوریتم‌هـای مـورد بررسـی دارای اهمیـت کمتـر از ۵۰درصد هستند. همچنین مشاهده مـی‌گـردد کـه الگـوریتم SPSA نسبت به سایر الگوریتم‌ها نتایج متفاوت‌تری را نشان می‌دهد.

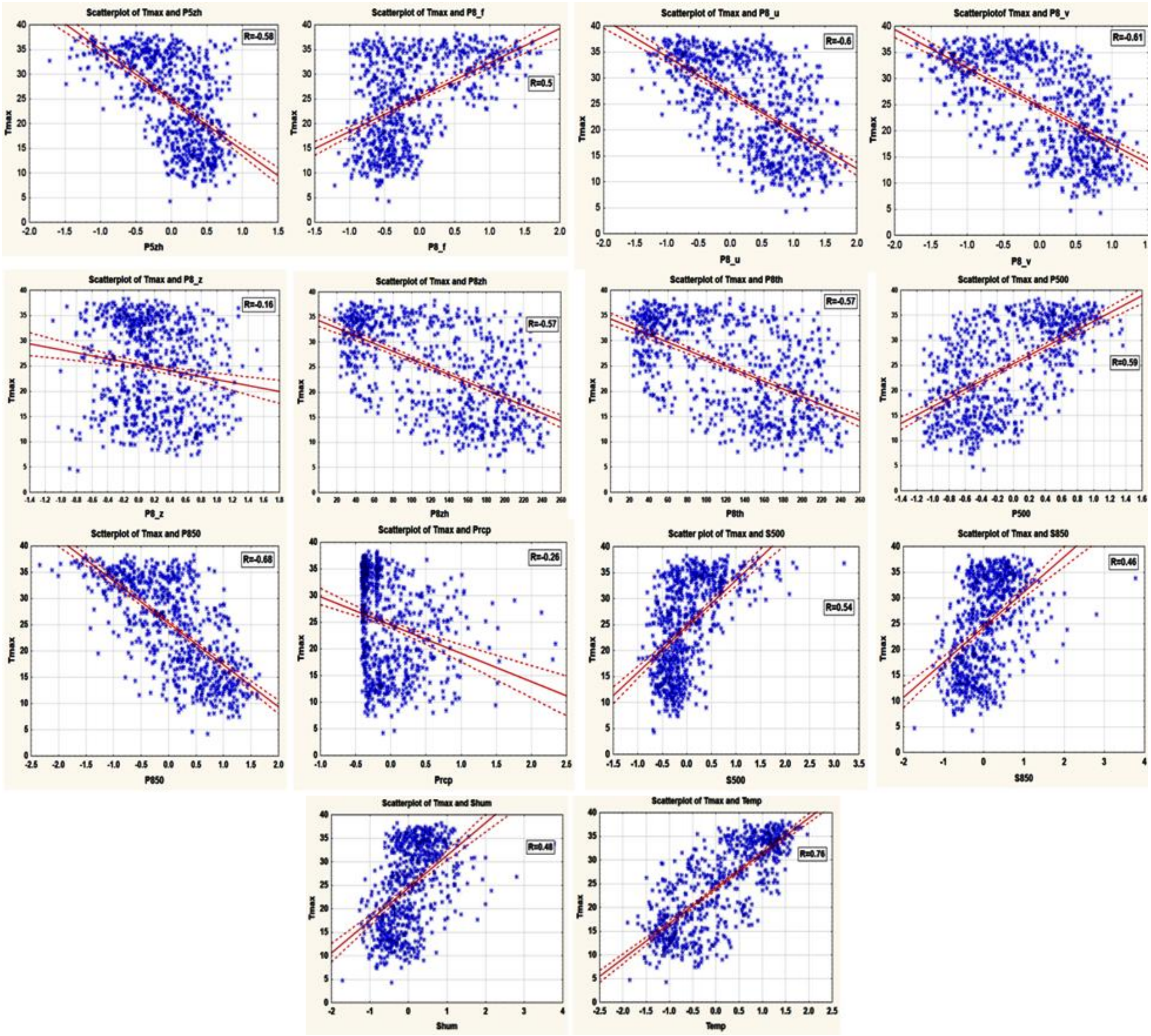


شکل ۲- بررسی تغییرات دمای حداکثر با ۱۲ پیش‌بینی‌کننده‌ی بزرگ مقیاس

(خط بهترین برازش و دامنه اطمینان و مقادیر همبستگی ارائه شده است)

Figure 2- Investigation of variation of maximum temperature with 12 large-scale predictors

(Best fit line and confidence intervals and correlation values are provided).



شکل ۳- بررسی تغییرات دما حداکثر با ۱۴ پیش‌بینی‌کننده‌ی بزرگ مقیاس

(خط بهترین برازش و دامنه اطمینان و مقادیر همبستگی ارائه شده است).

Figure 3- Investigation of variation of maximum temperature with 14 large-scale predictors

(Best fit line and confidence range and correlation values are provided).

برای بررسی توأمان میزان اهمیت مشخص شده در همه الگوریتم‌ها، میانگین میزان اهمیت در شکل ۵ نشان داده شده است. همچنین در این شکل خط اهمیت ۵۰ درصد نیز بعنوان متوسط اهمیت ترسیم شده است. نتایج نشان می‌دهد که پیش‌بینی‌کننده‌های ۱، ۳، ۴، ۶، ۹، ۱۵، ۱۶، ۱۸، ۱۹، ۲۰، ۲۱، ۲۶ دارای اهمیت بیشتر از ۵۰ درصد در ریزمقیاس نمایی دمای بیشینه در ایستگاه سینوپتیک بیرجند می‌باشند. بیشترین میزان اهمیت برای P1_v و کمترین مقدار آن در P5_u می‌باشد که مقادیر آن بترتیب ۷۳/۲٪ و ۱۵٪ بوده است.

نتایج ارزیابی الگوریتم‌های چهارگانه یادگیری ماشین در انتخاب پیش‌بینی‌کننده‌ها در شکل ۶ نشان داده شده است. هدف این الگوریتم‌ها ریزمقیاس‌نمایی دمای بیشینه در مقیاس ماهانه بوده است. بر اساس شاخص ارزیابی عملکرد نش- ساتکلیف نسبی، الگوریتم SPSA دارای عملکرد بسیار بالاتر از سایر الگوریتم‌ها بوده است بطوری که مقدار آن ۰/۶۸ بدست آمد در حالی که سایر الگوریتم‌ها مقادیر نزدیک به ۰/۳۷را تولید نموده‌اند. بر اساس شاخص کلینگ- گوپتا و بازده حجمی نیز نتایج الگوریتم SPSA دارای عملکرد بهتری از سه الگوریتم دیگر از خود نشان می‌دهد. همچنین بر اساس شاخص‌های ارزیابی بازده حجمی و نش-ساتکلیف نسبی، الگوریتم GBM از دو الگوریتم Ridge و LASSO در انتخاب پیش‌بینی‌کننده‌ها موفق‌تر بوده است.

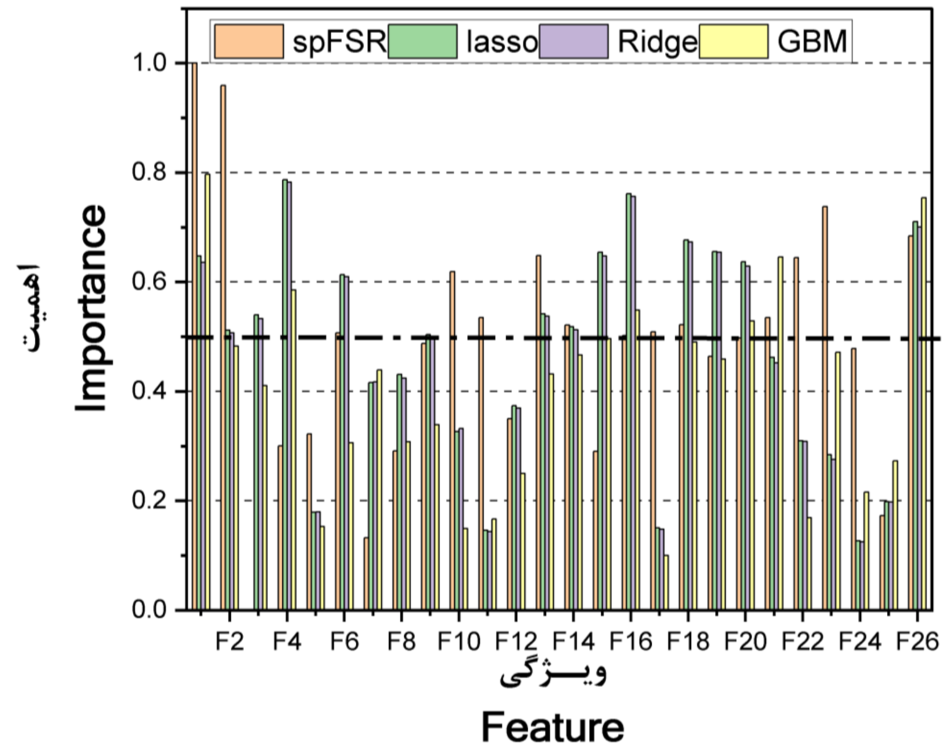
جدول ۲- بررسی همبستگی غیر خطی دمای بیشینه مشاهداتی و 26متغیر پیش‌بینی‌کننده

Table 2- Nonlinear correlation between observed maximum temperature and 26 predictor variable

همبستگی Correlation	متغیربزرگ مقیاس Large scale variable	ویژگی Property	همبستگی Correlation	متغیربزرگ مقیاس Large scale variable	ویژگی Property
0.8	میانگین فشار سطح دریا (mslp) Mean sea level pressure	F1	0.5	نیروی جریان هوا در ارتفاع 850hpa (p8_f) 850 hPa Wind speed	F14
0.54	سرعت جریان هوا ژئوستروفیک در نزدیک سطح (p1_f) 1000 hPa Wind speed	F2	0.61	مؤلفه سرعت مداری در ارتفاع 850hpa (p8_u) 850 hPa Zonal velocity	F15
0.49	مؤلفه سرعت مداری نزدیک سطح (p1_u) 1000 h Pa Zonal velocity	F3	0.61	مؤلفه سرعت نصف النهاری در ارتفاع 850hpa (p8_v) 850 h Pa Meridional velocity	F16
0.66	مؤلفه سرعت نصف النهاری نزدیک سطح (p1_v) 1000 hPa Meridional velocity	F4	0.16	تاوایی در ارتفاع 850hpa (p8_z) 850 hPa Vorticity	F17
0.31	تاوایی نزدیک سطح (p1_z) 1000 hPa Vorticity	F5	0.57	جهت باد در ارتفاع hPa۸۵۰ (p8th) 850 hPa Wind direction	F18
0.45	واگرایی نزدیک سطح (p1_th) 1000 hPa Wind direction	F6	0.59	واگرایی در ارتفاع 850hpa (p8zh) 850 hPa Divergence	F19
0.52	جهت باد نزدیک سطح (p1zh) 1000 hPa Divergence	F7	0.6	ارتفاع ژئوپتانسیل سطح 500hpa (p500) 500 hPa Geopotential height	F1\20
0.4	نیروی جریان هوا در ارتفاع 500hpa (p5_f) 500 hPa Wind speed	F8	0.69	ارتفاع ژئوپتانسیل سطح 850hpa (p850) 850 hPa Geopotential height	F21
0.42	مؤلفه سرعت مداری در ارتفاع 500hpa (p5_u) 500 hPa Zonal velocity	F9	0.26	(prcp) Precipitationبارش	F22
0.29	مؤلفه سرعت نصف النهاری در ارتفاع 500hpa (p5_v) 500 hPa Meridional velocity	F10	0.54	رطوبت نسبی در ارتفاع 500hpa (s500) 500 hPa Specific humidity	F23
0.36	تاوایی در ارتفاع 500hpa (P5_z) 500 hPa Vorticity	F11	0.46	رطوبت نسبی در ارتفاع 850hpa (s850) 850 hPa Specific humidity	F24
0.18	جهت باد در ارتفاع 500hpa (p5th) 500 h Pa Wind direction	F12	0.48	رطوبت ویژه در نزدیک سطح (shum) 1000 hPa Specific humidity	F25
0.58	واگرایی در ارتفاع 500hpa (p5zh) 500 hPa Divergence	F13	0.77	میانگین دما در ارتفاع ۲ متری (temp) Screen(2 m) air temperature	F26

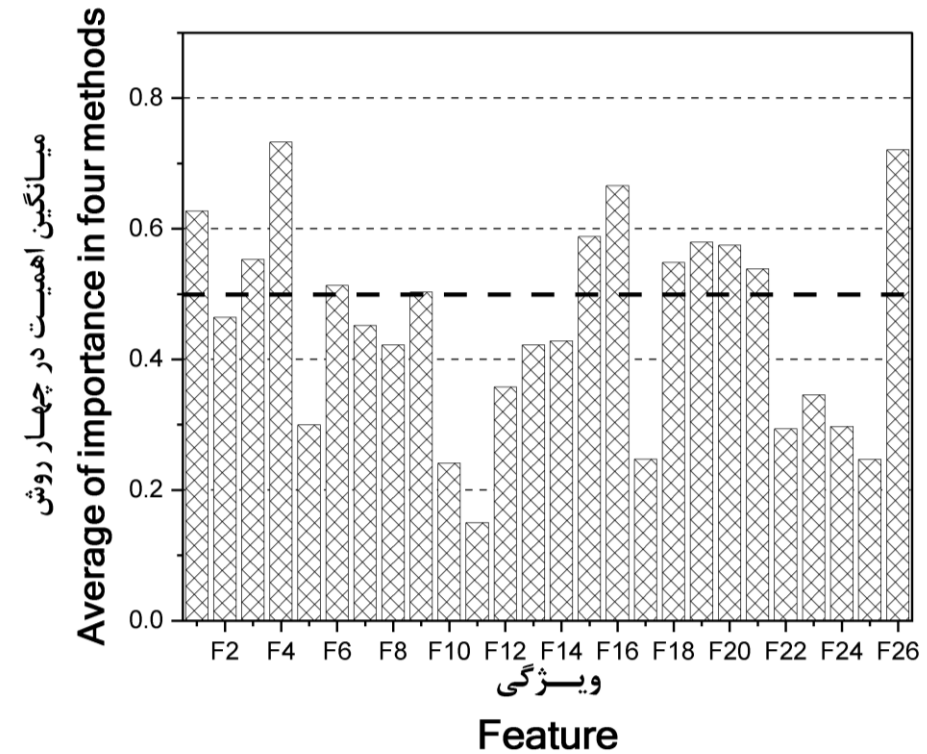
نمودار ویولین خروجی‌های الگوریتم‌های مختلف در کنار دمای بیشینه مشاهداتی در ایستگاه سینوپتیک بیرجند در شکل ۷ نشان داده شده است. این نمودار تأثیر انتخاب پیش‌بینی‌کننده‌ها در هر یک از الگوریتم‌ها در مرحله صحت‌سنجی نشان می‌دهد. طبق این نمودار مقادیر چارک‌های اول و سوم در داده‌های خروجی الگوریتم SPSA نسبت به سایر الگوریتم‌ها به داده‌های مشاهداتی نزدیک‌تر می‌باشند. پراکندگی داده‌های خروجی الگوریتم‌های LASSO و Ridge به داده‌های مشاهداتی نزدیک‌تر می‌باشد. پراکندگی خروجی الگوریتم های GBM نسبت به سایر الگوریتم‌ها کوچکتر می‌باشد. چگالی احتمال در محدوده چارک اول در الگوریتم‌های مورد استفاده مشابه داده‌های مشاهداتی بزرگتر از محدوده میانه و میانگین می‌باشد. در محدوده چارک سوم مقدار چگالی احتمال در الگوریتمSPSA و GBM مشابه داده‌های مشاهداتی می‌باشد و سایر الگوریتم‌ها چگالی احتمال متفاوت با داده‌های مشاهداتی پیش‌بینی نموده‌اند.

برای بررسی برابر بودن میانگین و واریانس خروجی ریزمقیاس شده توسط الگوریتم‌های مورد استفاده و داده‌های مشاهداتی در بخش صحت‌سنجی بترتیب از آزمون‌های آماری t-test و F-test استفاده گردید و در سطح معنی‌داری ۵٪ مورد تحلیل قرار گرفت (Hashmi *et al.*, 2011). نتایج این آزمون‌ها در جدول ۲ نشان داده شده است. نتایج آزمون مقایسه میانگین نشان می‌دهد که تنها میانگین در خروجی الگوریتم SPSA اختلاف معنی‌داری با میانگین داده‌های مشاهداتی در سطح ۵٪ ندارد. در سایر الگوریتم‌ها نتیجه آزمون نشان می‌دهد که فرض H_0 رد شده و لذا در سطح معنی‌داری ۵٪، اختلاف معنی‌دار بین خروجی سایر الگوریتم‌ها و میانگین داده مشاهداتی وجود دارد. لذا بر اساس آزمون بررسی تفاوت معنی‌داری بین میانگین خروجی الگوریتم‌های مورد بررسی و داده‌های مشاهداتی مشخص شد که الگوریتم SPSA نسبت به سایر الگوریتم‌ها کارایی بهتری دارد.



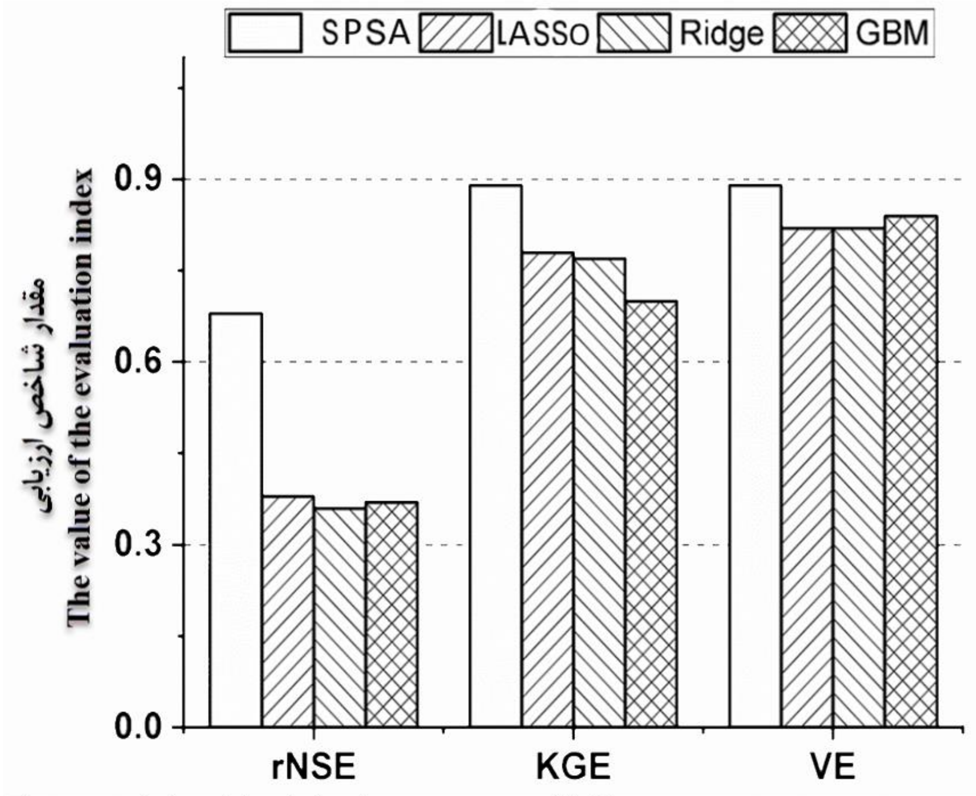
شکل ٤- نتایج بررسی اهمیت هریک از ۲٦ پیش‌بینی‌کننده بر اساس روش انتخابی

Figure 4- Importance results of each of the 26 predictors based on the selected algorithm



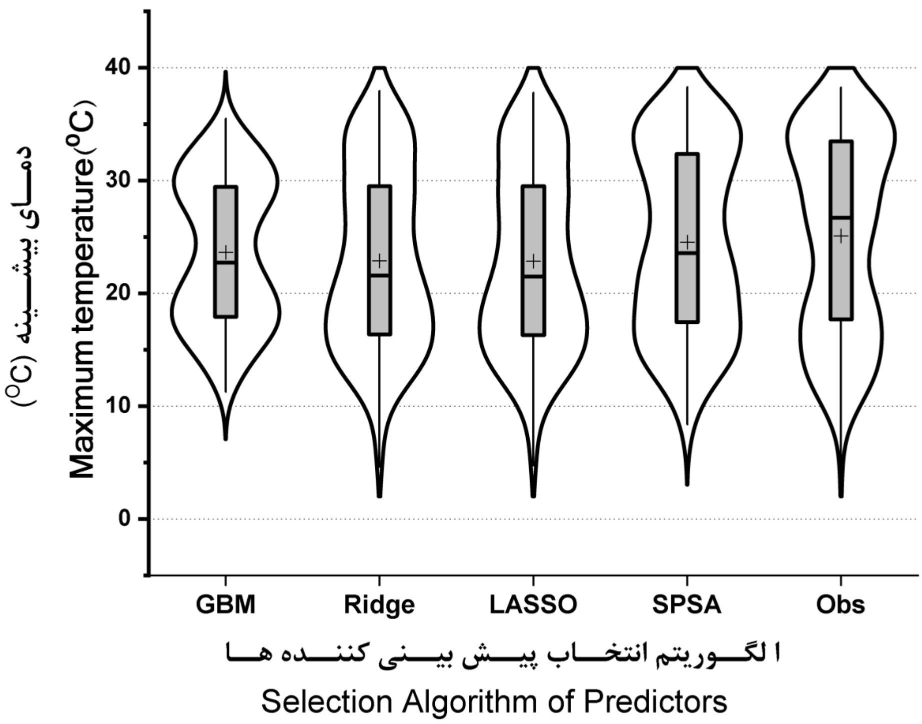
شکل ٥- بررسی میانگین اهمیت هر یک از ۲٦ پیش‌بینی‌کننده بزرگ مقیاس

Figure 5- Investigation of the average importance of each of the 26 large scale predictors



شکل ٦- مقادیر شاخص‌های ارزیابی برای روشهای مختلف انتخاب پیش بینی‌کننده

Figure 6- Values of evaluation criteria for different methods of predictor selection



شکل ۷- مقایسه نمودار باکس – پلات الگوریتم های مورد استفاده در مقایسه با دمای بیشینه مشاهداتی

Figure 7- Comparison of box-plot diagrams of algorithms used in comparison with the observed maximum temperature

جدول ۳- نتایج آزمون t-test و F-test روی خروجی الگوریتم‌های مورد استفاده در انتخاب پیش‌بینی‌کننده‌ها و درجه حرارت بیشینه مشاهداتی در ایستگاه سینوپتیک بیرجند در مرحله صحت‌سنجی

Table 3- Results of t-test and F-test on the output of algorithms used in selecting predictors and maximum observed temperature in Birjand synoptic station in validation stage

آزمون مقایسه آماره در داده‌های الگوریتم و مشاهداتی Statistical Comparison Test in Algorithm and Observational Data	SPSA	LASSO	Ridge	GBM
	P-value	P-value	P-value	P-value
(t-test) میانگین Average(t-test)	0.47	0.005*	0.006*	0.048*
(F-test) واریانس Variance(F-test)	0.31	0.19	0.15	1.3E-5*

آزمون بررسی واریانس خروجی الگوریتم‌های مورد بررسی و داده های مشاهداتی نشان داد که فرض H_0 رد شده و بنابراین تنها الگوریتم GBM دارای تفاوت معنی‌دار با واریانس داده‌های مشاهداتی در سطح ۵٪ می‌باشد. بنابراین بر اساس دو آزمون برابری میانگین و واریانس، الگوریتم SPSA دارای کارایی بالاتری نسبت به سایر الگوریتم‌ها در ریزمقیاس‌نمایی دمای بیشینه در ایستگاه سینوپتیک بیرجند می‌باشد.

نتیجه‌گیری

در این تحقیق برای انتخاب پیش‌بینی‌کننده‌های مهم برای ریزمقیاس نمائی آماری دمای بیشینه در ایستگاه بیرجند از چهار الگوریتم یادگیری ماشین شامل لاسو، ستیغی، GBM، SPSA استفاده گردید. داده‌های مورد استفاده در این تحقیق شامل متغیرهای جو بالا و نیز دمای بیشینه مشاهداتی در این ایستگاه بود که این الگوریتم‌ها برای انتخاب پیش‌بینی‌کننده‌های با اهمیت مورد استفاده قرار گرفتند و عملکرد این روش‌ها با شاخص‌های نش-ساتکلیف نسبی، کلینگ-گوپتا وبازده حجمی در بخش صحت‌سنجی مورد تحلیل قرار گرفت. همچنین میزان اهمیت هریک از متغیرهای ورودی جو بالا در ریزمقیاس‌نمایی دمای بیشینه نشان داده شد. ارتباط بین ۲۶ متغیر جو بالا و دمای بیشینه نشان داد که دو پیش‌بینی‌کننده (F1)mslp و (F26)Temp دارای بالاترین ضرایب همبستگی و P8_z دارای کمترین ضریب همبستگی با درجه حرارت بیشینه در ایستگاه دیدبانی بیرجند می‌باشند. اهمیت هر یک از ۲۶ پارامتر پیش-بینی‌کننده در ریز مقیاس‌نمایی دمای بیشینه با استفاده از الگوریتم های چهارگانه مورد مطالعه نشان داد که بیشترین میزان اهمیت برای P1_v(F4) و کمترین مقدار آن در P5_u(F4) می‌باشد که مقادیر آن بترتیب ۷۳/۲٪ و ۱۵٪ بوده است. مقایسه الگوریتم‌های مورد استفاده در این تحقیق برای داده‌های بخش صحت سنجی بر اساس شاخص ارزیابی عملکرد نش- ساتکلیف نسبی و کلینگ-گوپتا، نشان داد که الگوریتم SPSA دارای عملکرد بهتری از سایر الگوریتم‌ها درانتخاب پیش‌بینی‌کننده‌ها و به تبع آن ریزمقیاس‌نمایی دمای بیشینه می‌باشد. همچنین بررسی برابر بودن میانگین و واریانس خروجی ریزمقیاس شده توسط الگوریتم‌های مورد استفاده و داده‌های مشاهداتی در بخش صحت‌سنجی نشان دادکه الگوریتم SPSA دارای کارایی بالاتری نسبت به سایر الگوریتم‌ها درریزمقیاس‌نمایی دمای بیشینه در ایستگاه سینوپتیک بیرجند می‌باشد. بنابراین نتایج بررسی‌های چندگانه نشان داد که الگوریتم SPSA نسبت به سایر الگوریتم‌ها دارای توانمندی بالاتری در انتخاب پیش‌بینی‌کننده‌ها و ریزمقیاس‌نمایی دمای بیشینه در ایستگاه سینوپتیک بیرجند می‌باشد. همچنین با توجه به اینکه این تحقیق روی داده‌های یک ایستگاه صورت گرفته است، لذا بایستی نتایج آن برای داده های بیشتری مورد بررسی و تحلیل قرار گیرد و این می‌تواند بعنوان یک ایده در پژوهش‌های بعدی مورد استفاده قرار گیرد.

منابع

1. Aksakalli, V., & Malekipirbazari, M. (2016). Feature selection via binary simultaneous perturbation stochastic approximation, *Pattern Recognition Letters* 75: 41-47. https://doi.org/10.1016/j.patrec.2016.03.002.
2. Algin, R., Alkaya, A.F., & Agaoglu, M. (2022). *Performance of simultaneous perturbation stochastic approximation for feature selection*. In International Conference on Intelligent and Fuzzy Systems (pp. 348-354). Springer, Cham.
3. Babazadeh, H., Shamsnia, S.H., Bostani, F., Norozieghdam, A., & Khodakaramidwhkordi, D. (2012). Evaluation of drought, wet and prediction of Shiraz climatic parameters precipitation and temperature by using stochastic

methods. *Journal of Geography and Urban Planning 16*(41): 23-42. (In Persian)

4. Balling Jr, R.C., & Idso, S.B. (1990). Effects of greenhouse warming on maximum summer temperatures. *Agricultural and Forest Meteorology 53*(1-2): 143-147.
5. Chen, H., Xu, C.Y., & Guo, S.L. (2012). Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology* 434: 36–45. https://doi.org/10.1016/j.jhydrol.2012.02.040.
6. Diamantopoulu, M.J., Georgiou, P.E., & Papamichial, D.M. (2010). Evaluation of artificial neural network in estimating reference evapotranspiration with minimal meteorological data. *Global Nest Journal 13*(1): 18-27.
7. Fatahi, M.H., Bamdad, A., & Rahimikhob, A. (2012). Application of association rules to monitor rainfall and drought events using sea surface temperature (Case study: Khozestan). *Journal of Water Resource Engineering* 109-118. (In Persian with English abstract)
8. Hashmi, M.Z., Shamseldin, A.Y., & Melville, B.W. (2011). Comparison of SDSM and LARS-WG for simulation and downscaling of extreme precipitation events in a watershed. *Stochastic Environmental Research and Risk Assessment 25*(4): 475-484.
9. He, R.R., Chen, Y., Huang, Q., & Kang, Y. (2019). LASSO as a tool for downscaling summer rainfall over the Yangtze River valley. *Journal of Hydrology 64*(1): 92–104. https://doi.org/10.1080/02626667.2019.1570210.
10. Hessami, M., Gachon, P., Ouarda, T., & St-Hilaire, A. (2008). Automated regression-based statistical downscaling tool. *Environ Model Software* 23: 813–834. (In Persian)
11. Jafarzadeh, A., Pourreza-Bilondi, M., Khashei Siuki, A., & Ramezani Moghadam, J. (2021). Examination of various feature selection approaches for daily precipitation downscaling in different climates. *Water Resources Management 35*(2): 407-427.
12. Kharin, V.V., & Zwiers, F.W. (2000). Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM. *Journal of Climate 13*(21): 3760-3788. https://doi.org/10.1175/1520-0442(2000)013<3760:CITEIA>2.0.CO;2.
13. Meenu, R., Rehana, S., & Mujumdar, PP. (2013). Assessment of hydrologic impacts of climate change in Tunga-Bhadra River basin, India with HEC-HMS and SDSM. *Hydrological Process 27*(11): 1572–1589. https://doi.org/10.1002/hyp.9220.
14. Muthukrishnan, R., & Rohini, R. (2016). *LASSO: A feature selection technique in predictive modeling for machine learning*. In 2016 IEEE international conference on advances in computer applications (ICACA) (pp. 18-20). IEEE.
15. Nasseri, M., & Zahraie, B. (2013). Performance assessment of different data mining methods in statistical downscaling of daily precipitation. *Journal of Hydrology* 492: 1–14. https://doi.org/10.1016/j.jhydrol.2013.04.017.
16. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* 7: 21. https://doi.org/10.3389/fnbot.2013.00021.
17. Nazeri Tahroudi, M., Amirabadizadeh, M., & Zaineli, M.J. (2017). Investigating artificial intelligence and regression methods in simulating daily temperature values. *Meteorology and Atmospheric Sciences 1*(1): 65-76. (In Persian)
18. Niknam, F. (2013). *Climatic data mining to present a climate forecasting model in Isfahan Province*. University of Shiraz.
19. Omidvar, K., Shafii, Sh., Taghizadeh, Z., & Alipur, M. (2015). Efficient evaluation of decision tree model in Kermanshah Synoptic station rainfall forecast. *Journal of Applied Geosciences Research 14*(34): 89-110. (In Persian)
20. Pal, M., & Deswal, S. (2009). M5 model tree based modeling of reference evapotranspiration. *Hydrologic Process* 23: 1437-1443. https://doi.org/10.1002/hyp.7266.
21. Panahi, M., & Mirshahi, S.H. (2016). Assessment among two data mining algorithms CART and CHAID in forecast air temperature of the Synoptic station of Arak. *Journal of Environmental Science 13*(4): 52-58. (In Persian)
22. Salahi, B., & Fateminiya, F.S. (2017). Forecasting frost changes in the city of Kashan based on the simulation of general atmospheric circulation model. *Journal of Geography and Environmental Planning 28*(3): 20-36. (In Persian)
23. Sfandiari, F., Hosseini, S.H., Azadimobaraki, M., & Hejazizadeh, Z. (2010). Predict the average monthly temperature in Sanandaj station using the model (MLP) MLP Network, *Journal of Iran Geographic 8*(27): 45-65. (In Persian)
24. Troncoso, A., Salcedo_Sanz, S., Casanova_ Mateo, C., Riquelme, J.C., & Prieto, L. (2015). Local model based regression trees for very short-term wind speed prediction. *Renewable Energy* 81: 589-598.
25. Zhang, X., Yan, X., & Chen, Z. (2016). Reconstructed regional mean climate with Bayesian model averaging: a case study for temperature reconstruction in the Yunnan–Guizhou plateau, China. *Journal of Climate 29*(14): 5355–5361. https://doi.org/10.1175/JCLI-D-15-0603.1.